DOCUMENT RESUME

ED 367 669 TM 021 106

AUTHOR Beasley, T. Mark; Leitner, Dennis W.
TITLE The p-Problem with Stepwise Multiple Regression.
PUB DATE Feb 94
NOTE 16p.; Paper presented at the Annual Meeting of the Eastern Educational Research Association (17th, February 12, 1994).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS *Algorithms; Computer Simulation; Correlation; *Error of Measurement; Matrices; Monte Carlo Methods; Prediction; *Regression (Statistics); Research Problems; *Statistical Analysis; *Statistical Significance; Test Interpretation; Validity
IDENTIFIERS Bonferroni Procedure; Eigenvalues; Power (Statistics); P Values; Research Replication; *Stepwise Regression; *Type I Errors

ABSTRACT

The use of stepwise regression has been criticized for both interpretive misunderstandings and statistical aberrations. A major statistical problem with stepwise regression and other procedures that involve multiple significance tests is the inflation of the Type I error rate. General approaches to control the family-wise error rate such as the Bonferroni correction are based on the assumptions of independent tests which typically reduces power. Because correlated predictors is a more realistic situation, other algorithms based on the average correlation in the predictor matrix have been proposed. The present study proposes an algorithm based on the maximum eigenvalue and the determinant of the predictor matrix for controlling the family-wise Type I error rate for multiple corrected tests. A Monte Carlo simulation with 5,000 replications was performed to demonstrate the effectiveness of the proposed algorithm. Three tables and one figure are included. (Contains 19 references.) (Author)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

TM



ED 367 669

# The *p*-Problem with Stepwise Multiple Regression

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

T. Mark Beasley
St. John's University

Dennis W. Leitner
Southern Illinois University

"PERMISSION TO REPRODUCE THIS MATERIAL IN **OTHER THAN PAPER COPY** HAS BEEN GRANTED BY
T. Mark Beasley
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper presented at the meeting of the
Eastern Educational Research Association
February 12, 1994

Direct all correspondence to:
T. Mark Beasley
DAIL, School of Education
St. John's University
Jamaica, NY 11439
beasleyt@sjuvm

TM021106



BEST COPY AVAILABLE

Abstract

The use of stepwise regression has been criticized for both interpretive misunderstandings and statistical aberrations. A major statistical problem with stepwise regression and other procedures that involve multiple significance tests is the inflation of the Type I error rate. General approaches to control the family-wise error rate such as the Bonferroni correction are based on the assumptions of independent tests which typically reduces power. Because correlated predictors is a more realistic situation, other algorithms based on the average correlation in the predictor matrix have been proposed. The present study proposes an algorithm based on the maximum eigenvalue and the determinant of the predictor matrix for controlling the family-wise Type I error rate for multiple, correlated tests. A Monte Carlo simulation with 5,000 replications was performed to demonstrate the effectiveness of the proposed algorithm.

For several years now, applied statisticians (e.g., Pope & Webster, 1972; Thompson, 1988; Wilkinson, 1979) have been calling attention to the abuses of stepwise regression in its common use by less statistically sophisticated researchers. But the theses and dissertations continue to step (unwisely) across our desks and articles with many of these same problems continue to appear in print. It is hoped that elaborating the limitations of stepwise regression will bring about its more constrained use. General methodological alternatives are available and a procedure to control Type I errors in stepwise regression is proposed.

Three statistical procedures are used under the rubric of stepwise regression: Forward selection; backward elimination; and true stepwise (Draper & Smith, 1981). The *forward selection* procedure forms a model from the set of $k$ dependent variables by first selecting the single best predictor. The second best predictor is then chosen by the criteria of strongest contribution to the prediction of $Y$, while controlling for the effects of the first predictor entered. Thus, the first step involves $k$ simultaneous tests of zero-order correlations, while the second step involves ($k$ - 1) simultaneous tests of first-order semi-partial correlations (Aitkin, 1974). The process continues so that at each step the variable selected for inclusion significantly increases the prediction of $Y$. The *backward elimination* procedure begins with a model containing all $k$ variables and then at each step a variable is eliminated if its removal results in the smallest reduction in the model's effectiveness without significantly reducing the model $R^2$. The *true stepwise* procedure is a variant of the forward selection technique. It differs from forward selection in that at each step, a variable that has been entered in the forward manner can be removed if a partial $F$-test shows that variable to be a nonsignificant predictor.

The use of the various stepwise regression procedures has been criticized for many interpretative misuses and statistical aberrations. First, researchers often interpret the final solution of a reduced set of $g$ predictors as being the best subset of predictors overall and of that size. Also, there is a tendency to confuse the order of entry and variable importance (Welge, 1990). Furthermore, the true stepwise procedure, as well as the backward elimination method, suffer from the use of the largest partial $F$ as a test of a potential entry variable which is not in the regression at that stage. The correct null sampling distribution for this test in not the ordinary $F$ distribution, but is a

partial *F* distribution which is very difficult to obtain. (Draper & Smith, 1981). Moreover, researchers proceed to test each stage in a stepwise regression as if the partial *F* distribution does not exist and as if the test at that step is the only test that has or will occur. Furthermore, the degrees of freedom (*df*) used for these tests are often incorrect. For example in the forward selection procedure the *df*'s used for the first step, a test of zero-order correlations, is $(N - 2)$, while $(N - k - 1)$ would be more appropriate. These consideration, in general, tend to inflate the probability of at least one Type I error, that is, the probability of forming an erroneous model. Another interpretative problem arises when two or more predictor variables are highly correlated. In such situations, there is a strong probability that one of the variables will absorb the majority of the other variables' prediction power are therefore cause their exclusion from subsequent models. Not only does a set of correlated predictors lead to potential substantive misinterpretations, it also makes estimating the probability of a Type I error more complex. Thus, due to multiple tests, incorrect *df*'s, misunderstood partial *F* tests, and correlated predictor variables, the correct family-wise Type I error rate in stepwise regression is difficult to determine. To compound this problem, the *p*-value associated with each variable entered stepwise into a regression equation (except for the final step) is incorrect in many canned statistical packages. Thus, the focus of this paper is the erroneous *p*-value problem in stepwise regression.

The true family-wise Type I error rate ($\alpha_T$) for $k$ independent tests with the same alpha level is defined by the following equation:

$$\alpha_T = 1 - (1 - \alpha_i)^k. \qquad (1)$$

The forward selection method in stepwise regression conducts no less than $k$ simultaneous tests of significance as if multiple tests are not performed. That is, the first predictor is selected by the largest zero-order correlation of all $k$ variables without consideration for the number of tests being conducted. Assuming independent predictors, the probability of a Type I error on the first step is equal to (1), which in the null case distributes evenly across predictor variables. If a researcher wished to adjust $\alpha_T$ so that the Type I error rate returned to the nominal alpha ($\alpha_i$) then he could adjust $\alpha_i$ by the following formula:

$$\alpha_{adj} = 1 - (1 - \alpha_i)^{1/k}. \qquad (2)$$

If the $k$ predictors were all perfectly correlated, however, the Type I error rate would equal alpha (i.e., $\alpha_T = \alpha_i$). In the more realistic situation of

correlated predictors, the solution for the correct Type I error rate is considerably more complex and requires the integration of the correlated *F* distribution (Pope & Webster, 1972). Furthermore, only limited tables of critical values are available (Games, 1977), while a few Monte Carlo approximations based on averaged correlations have been proposed (i.e., Krishnaiah & Armitage, 1965; Pohlmann, 1979). The appropriate Type I error rate, however, may approach (1) with *b* orthogonal factors rather than these algorithms based on the average correlation of *k* predictors. To elaborate, one possible approach to the *p*-value problem would be to perform a principal component analysis (PCA) on the predictor correlation matrix and extract *b* orthogonal components. In fact, it can easily be shown that such a linear transformation will not affect the model $R^2$. That is, if all *k* variables or if *k* components extracted from the predictor matrix are used as separate models to predict a criterion variable, *Y*, then both would have the same full model $R^2$. The Type I error rate when using the *k* orthogonal components, however, will equal (1) for the first step of a forward selection stepwise regression. Thus, decomposing the set of *k* predictors into *b* orthogonal components and modifying algorithms for correlated predictors may provide a better approximation of the family-wise Type I error rate. However, this indicates a relationship between the transformation matrix and the family-wise Type I error rate. Specifically, it is proposed that the maximum eigenvalue ($\lambda_{max}$) from an unrotated principal components analysis and the determinant, |**P**|, of the predictor correlation matrix, **P**, is related to the proportion of Type I errors on the first-step, which defines the probability of forming an erroneous model of one or more predictors.

The eigenvalues of a correlation matrix, **R**, are commonly used as indices of the number of factors that underlie a correlation matrix (e.g., Kaiser, 1970). Furthermore, the maximum eigenvalue provides an index for the proportion of variance accounted for by the largest principal component, the average correlation of **R**, and the number of underlying factors (Tatsuoka, 1988). The determinant of a correlation matrix has been used in establishing the independence of variables in PCA (Nagarsenker, 1976). Since the determinant of the covariance matrix, **C**, gives the generalized variance (Tatsuoka, 1988),

and the determinant of the correlation matrix, **R**, is equal to |**C**| divided by the determinant of the diagonal variance matrix **V**,

$$|\mathbf{R}| = \frac{|\mathbf{C}|}{|\mathbf{V}|}, \tag{3}$$

then it follows that the generalized proportion of variance in **R**, that is the generalized $\mathbf{R}^2$, is equal to:

$$\mathbf{R}^2 = 1 - \frac{|\mathbf{C}|}{|\mathbf{V}|} = 1 - |\mathbf{R}|. \tag{4}$$

Thus, in combination $\lambda_{max}$ and |**P**| provide rather unique information about the inter-correlation of the predictor matrix, **P**. Specifically in PCA, $\lambda_{max}$ divided by $k$ gives the proportion of variance in **R** accounted for by the first and largest principal component. However, $\lambda_{max}$ is known to always be greater than one even in random data matrices (Horn, 1965). In fact, when the variables are independent and all off-diagonal elements in **R** are zero then $\lambda_{max}$ equals one,

$$\lim_{\mathbf{R} \to 0} \lambda_{max} = 1. \tag{5}$$

Therefore, subtracting one from $\lambda_{max}$ and dividing by $k$ would provide a corrected proportion of variance for the largest principal component

$$\frac{\lambda_{max} - 1}{k} \tag{6}$$

Also, if the variables are independent then |**R**| equals one,

$$\lim_{\mathbf{R} \to 0} |\mathbf{R}| = 1. \tag{7}$$

Although it is left undefined, one can imagine that if all predictor variables were perfectly correlated, then $\lambda_{max}$ would equal $k$. That is, the limit of $\lambda_{max}$ as all the elements of **P** approach unity is $k$:

$$\lim_{\mathbf{P} \to 1} \lambda_{max} = k. \tag{8}$$

Furthermore, since the product of the eigenvalues must equal the determinant, then under the same conditions specified for (8), the limit of |**P**| equals zero as $\lambda_{max}$ approaches $k$ :

$$\lim_{\mathbf{P} \to 1} |\mathbf{P}| = 0 \quad \text{and} \quad \lim_{\lambda_{max} \to k} |\mathbf{P}| = 0. \tag{9}$$

Given conditions (8) and (9), there is only one "true" variable and the *family size*, $c$, should be equal to one, which can be described as

$$c = k - (k - 1) \tag{10}$$

Thus, if we multiply (6) times $(k - 1)$ we get the proportion of $(k - 1)$ that should be subtracted from $k$; however, $c$ also depends on the correlations in **P**

whose generalized estimate comes from |P|, and thus, $(k - 1)$ should be multiplied by (6) and (4). Therefore, $c$ can be estimated by:

$$c = k - \frac{(k - 1)(\lambda_{max} - 1)(1 - |\mathbf{P}|)}{k} . \quad (11)$$

Thus under the conditions set in (5), (7), (8), and (9), when the inter-correlation of the predictor variables approach unity, (11) approaches one. Also if the $k$ predictors are independent then (11) equals $k$. Therefore, if a researcher can use $k$, $\lambda_{max}$ ,and |P| to estimate the independence of the predictors or family size of **P** with $c$ , then (11) could be used to adjust $\alpha_T$ so that it approximates the nominal alpha,

$$\alpha_{adj} = 1 - (1 - \alpha_i)^{1/c} . \quad (12)$$

Thus in the present study, a Monte Carlo simulation of a forward-selection stepwise procedure with no expected correlation between the dependent variable, $Y$, and the $k$ predictors was used to estimate the correct $p$-values for $k$ = 2, 3, 4, 5, 7, and 10 correlated variables under various inter-predictor correlation matrices. From these results, a general formulation using $k$, $\lambda_{max}$ and |P| to estimate $c$ was proposed (11) and tested using (12) to determine whether it was useful in controlling Type I errors.

## Method

### *Simulation Procedure*

A Monte Carlo program was written in SAS/IML (SAS Institute, 1990) to simulate the forward selection process of stepwise regression. Initially, a predictor matrix, **X**, of dimensions $n$ cases (rows) and $k$ variables (columns) was randomly generated. Then by using the fundamental postulate of PCA (Tatsuoka, 1988) and a method described by Kaiser and Dickman (1962), a $k$ x $k$ matrix the principal component coefficients, **F**, was derived from **P** and pre-multiplied by the transpose of **X** to create a transformed data matrix $\mathbf{Z}_p$ with the particular configuration of correlations that simulates **P**.

$$\mathbf{Z}_p = \mathbf{F}\,\mathbf{X}^t \quad (13)$$

The number of predictors was manipulated from $k$ = 2, 3, 4, 5, 7, and 10. The number of cases was held constant at a fairly large number of $n$ = 200 in order to avoid extreme shrinkage of $R^2$ (Harris, 1975) and to reduce the residual error in **P** as it simulates the correlation matrix described by **F**. Then a dependent variable vector, **Y**, was randomly generated and concatenated with the transpose $\mathbf{Z}_p$ to form **M**. Thus, although there was correlation among the $k$ variables in **P**, there was no expected correlation between the predictor

variables and $Y$. This process was replicated 5,000 times. Since an infinite number of correlation matrices can be simulated, various combinations of $\lambda_{max}$ and |P| were used for each level of $k$. Tables 1, 2, and 3 in the results section reference the values of $\lambda_{max}$ and |P| that were imposed on **X**.

*Test Procedures*

In the true stepwise and forward selection procedures, if an erroneous model is to be formed (i.e., Type I error committed) then it will occur on the first step. Furthermore most packaged stepwise programs (i.e., SAS STEPWISE, SPSS REGRESSION) perform the first entry with $(N - 2)$ *df*'s. Therefore, the maximum correlation in the first column of **M** was tested. If the calculated $F(1, 180)$ exceeded the $\alpha = .05$ level then it was counted as a Type I error. The number of rejections was divided by the 5,000 replications which served as empirical *p*-values and estimates of $\alpha_T$. Thus, the results of this procedure were used to help estimate $c$. If $\alpha_T$ and $\alpha_i$ are known then $c$ can be solved as follows:

$$c = \frac{\ln(1 - \alpha_T)}{\ln(1 - \alpha_i)}. \qquad (14)$$

These values of $k$, $\lambda_{max}$, and |P| using the formula described in (11) were regressed on $c$ from (14) to demonstrate the goodness of fit. Furthermore, the effectiveness of (11) was assessed by using these estimates of $c$ in (12) to set a more stringent $\alpha_i$ in order to control the Type I error rate.

## Results

Using the expected values of $k$, $\lambda_{max}$, and |P|, the family size estimates of $c$ from (11) were regressed on the empirical values of $c$ derived from the proportion of rejections at the $\alpha = .05$ level of significance during the 5,000 replications and (14). Thus, the following model was tested

$$c_{emp} = b_1 k + b_2 \frac{(k - 1)(\lambda_{max} - 1)(1 - |\mathbf{P}|)}{k} + e. \qquad (15)$$

with the intercept restricted to zero and the coefficients $b1$ and $b_2$ restricted to one. The model $R^2$ with these restriction was 0.9858. Figure 1 (upper panel) shows a scatter plot of this analysis with different element for each value of $k$. The model $R^2$ when using Pohlmann's formula based on averaged squared correlations was 0.9147. A scatter plot of that regression is shown in the lower panel of Figure 1. The diagonals on each panel represent a perfect fit of the expected and empirical values of family size. Using a dependent *t*-test for correlations, these models were found to be significantly different,

$t(67) = 9.70$, $p < .001$.. Thus, Figure 1 also shows that many more estimates of family size, $c$, deviate from the perfect fit diagonal for the Pohlmann's average squared correlation estimate of $c$ as compared to (11).

Tables 1, 2, and 3 show the expected values of the average squared correlation within the predictor matrix, **P**, the maximum eigenvalue ($\lambda_{max}$), and the determinant of **P**, |**P**|. These tables also show the empirical values of the family-wise Type I error rate ($p$-values), the estimated value of family size, $c$ from (14), and the corrected $p$-values after controlling Type I errors with (11) and (12).

As expected, when number of variables increases from 2 to 10 the family-wise Type I error rate also increases, as can be seen by looking across Tables 1, 2, and 3. One can also see by looking within any Table that as the expected average squared correlation increases the Type I error rate decreases. Yet, some matrices with the same average $r^2$ have different values for $\lambda_{max}$, and |**P**| and more importantly different empirical proportions of Type I errors. This is most notable in Table 3 with $k = 10$ and it is important to note that the corrected $p$-values using (11) and (12) are reasonably close to the nominal alpha of .05 regardless of these discrepancies. Corrections based on average $r^2$s (i.e., Pohlmann, 1979), however, would correct these discrepant configurations in the same manner. Thus, logically as well as statistically, the correction formula (11) provides a better adjustment for controlling Type I errors for multiple, correlated tests.

## Discussion

The behavioral science literature is replete with "significant" findings that fail the ultimate test of replication (Rosnow & Rosenthal, 1989). One explanation for this conundrum lies in the family-wise Type I error rate that increases when stepwise regression or other multiple comparison procedures are used. Faced with the problem of multiple tests that may be correlated, the researcher should take some correction action. Possible approaches to this problem include:

a). Prior to performing a stepwise regression, conduct an omnibus test with all potential predictors in the model.

b). When searching for a significant subset of predictors, use stepwise methods with backward elimination

c). When searching for a reduced subset of predictors through stepwise methods, perform a PCA and extract orthogonal components and use

(1) to correct the family-wise Type I error rate.

d). In any multiple test situation, use one of several simultaneous inference tests (e.g., Games, 19977; Schafer, 1992; Schafer & Macready, 1975) to control Type I errors.

e). Use the Bonferroni inequality, however, one may over-correct the probability of a Type I error and lose power.

f). Use the algorithm suggested here, (11) and (14), if multiple, correlated test are being performed.

It should be noted that there are practically an infinite number of configurations a correlation matrix can assume; therefore, there is no way to exhaust those possibilities. Therefore, these findings are limited to the specific correlation matrices simulated. Thus, although extensive replications of this study are needed to assume the generality of these findings, it is not unreasonable to assume that the proposed algorithm (11) will work in other situations.

Although the family-wise Type I error correction suggested here has been framed in terms of stepwise regression, there is no reason for its exclusion from other situations that involve a single dependent variable and multiple tests that are correlated. For example, a set of nonorthogonal contrasts for an analysis of variance, although based on coded vectors for means have correlations coefficients associated with them. Therefore, a matrix of correlations among contrasts could be analyzed with (11). In conclusion, the suggested algorithm shows adequate control of the family-wise Type I error rate and is based on more complete information than estimates based simply on the average squared correlation. Yet, in the results the suggested correction sometimes deviated from the nominal alpha. Thus, further investigation will focus on manipulating sample sizes and using a shrinkage correction for the determinant of the predictor correlation matrix.

## References

Aitkin, M. A. (1974). Simultaneous inference and the choice of variable subsets in multiple regression. *Technometrics, 16*, 221-27.

Draper, N., & Smith, H. (1981). *Applied regression analysis* (2nd ed.). New York: Wiley.

Games, P. A. (1977). An improved table for simultaneous control on *g* contrasts. *Journal of the American Statistical Association, 72*, 531-534.

Harris, R. (1975). *A primer of multivariate statistics*. New York: Academic Press.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*, 179-185.

Kaiser, H. F. (1970). A second generation Little Jiffy. *Psychometrika, 35*, 401-415.

Kaiser, H. F., & Dickman, K. (1962). Sample and population score matrices and sample correlation matrices from an arbitrary population correlation matrix. *Psychometrika, 27*, 179-182.

Krishnaiah R. K. & Armitage, R. B. (1965). *Probability integrals of the multivariate F distribution, with Tables and applications*. (Report No. ARL 65-236). Wright-Patterson AFB, OH: U. S. Air Force.

Nagarsenker, B. N. (1976). The distribution of the determinant of correlation matrix useful in principal components analysis. *Communications in Statistics: Simulation and Computation, B5*, 1-13.

Pohlmann, J. T. (1979). Controlling the Type I error rate in stepwise regression analysis. *Multiple Linear Regression Viewpoints, 10*, 46-60.

Pope, P. T., & Webster, J. T. (1972). The use of an *F*-statistic in stepwise regression procedures. *Technometrics, 14*, 327-340.

Rosnow, R. L. &, Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44*, 1276-1284.

SAS Institute (1990). *SAS/IML user's guide* (Release 6.04). Cary, NC: Author.

Schafer, W. D. (1992). Simultaneous inference options for statistical decision making. *Measurement and Evaluation in Counseling and Development, 25*, 98-101.

Schafer, W. D., & Macready, G. B. (1975). A modification of the Bonferroni procedure on contrasts which are grouped into internally independent sets. *Biometrics, 31*, 227-228.

Thompson, B. (November, 1988). *Common methodology mistakes in dissertations: Improving dissertation quality*. Paper presented at the meeting of the Mid-South Educational Research Association. Louisville, KY.

Tatsuoka, M. M. (1988). *Multivariate analysis: Techniques for educational and psychological research* (2nd ed.). New York: Macmillan.

Welge, P. (January, 1990). *Three reasons why stepwise regression methods should not be used by researchers*. Paper presented at the meeting of the Southwest Educational Research Association. Austin, TX.

Wilkinson, L. (1979). Tests of significance in stepwise regression. *Psychological Bulletin, 86*, 168-174.

*Table 1.*
Expected values for the average $r^2$, maximum eigenvalue ($\lambda_{max}$) and determinant (|P|) with empirical Type I error rate, estimated $c$ from (14) and corrected Type I error rate (11) for $k$ = 2 and 3 predictors.

| $E(\bar{r}^2)$ | $E(\lambda_{max})$ | E( \|P\| ) | $k = 2$ *p*-value | $c$ (14) | Corrected *p*-value |
|---|---|---|---|---|---|
| 0.00 | 1.0000* | 1.0000 | 0.0962 | 1.9719 | 0.0506 |
| 0.01 | 1.1000* | 0.9900 | 0.0998 | 2.0497 | 0.0518 |
| 0.09 | 1.3000* | 0.9100 | 0.0934 | 1.9116 | 0.0496 |
| 0.25 | 1.5000* | 0.7500 | 0.0892 | 1.8215 | 0.0510 |
| 0.49 | 1.7000* | 0.5100 | 0.0896 | 1.8301 | 0.0532 |
| 0.64 | 1.8000* | 0.3600 | 0.0760 | 1.5410 | 0.0454 |
| 0.81 | 1.9000* | 0.1900 | 0.0714 | 1.4442 | 0.0438 |
| | | | $k = 3$ | | |
| 0.00 | 1.0000* | 1.0000 | 0.1474 | 3.1089 | 0.0502 |
| 0.09 | 1.5695 | 0.7609 | 0.1296 | 2.7061 | 0.0496 |
| | 1.5984 | 0.7826 | 0.1394 | 2.9268 | 0.0494 |
| | 1.6000* | 0.7840 | 0.1300 | 2.7150 | 0.0462 |
| 0.25 | 1.8922 | 0.2910 | 0.1120 | 2.3158 | 0.0452 |
| | 1.9860 | 0.4692 | 0.1286 | 2.6837 | 0.0516 |
| | 2.0000* | 0.5000 | 0.1250 | 2.6033 | 0.0540 |
| 0.49 | 2.3658 | 0.0700 | 0.1052 | 2.1670 | 0.0532 |
| | 2.3986 | 0.2531 | 0.1090 | 2.2500 | 0.0504 |
| | 2.4000* | 0.2160 | 0.1122 | 2.3202 | 0.0492 |
| 0.64 | 2.5885 | 0.0384 | 0.0910 | 1.8601 | 0.0462 |
| | 2.6000* | 0.1040 | 0.1004 | 2.0627 | 0.0510 |

*Note.* * indicates that all correlations in **P** are equal.

*Table 2.*
Expected values for the average $r^2$, maximum eigenvalue ($\lambda_{max}$) and determinant (|P|) with empirical Type I error rate, estimated $c$ from (14) and corrected Type I error rate (11) for $k$ = 4 and 5 predictors.

| $E(\bar{r}^2)$ | $E(\lambda_{max})$ | E( \|P\| ) | $k = 4$ *p*-value | $c$ (14) | Corrected *p*-value |
|---|---|---|---|---|---|
| 0.00 | 1.0000* | 1.0000 | 0.1870 | 4.0361 | 0.0494 |
| 0.09 | 1.7926 | 0.5832 | 0.1744 | 3.7363 | 0.0470 |
| | 1.8016 | 0.5439 | 0.1706 | 3.6467 | 0.0474 |
| | 1.8964 | 0.6481 | 0.1746 | 3.7410 | 0.0478 |
| | 1.9000* | 0.6517 | 0.1790 | 3.8452 | 0.0552 |
| 0.25 | 2.2670 | 0.1188 | 0.1548 | 3.2788 | 0.0512 |
| | 2.4150 | 0.1042 | 0.1508 | 3.1868 | 0.0528 |
| | 2.4995 | 0.3019 | 0.1542 | 3.2650 | 0.0582 |
| | 2.5000* | 0.3125 | 0.1650 | 3.5155 | 0.0572 |
| 0.64 | 3.3696 | 0.0011 | 0.1042 | 2.1453 | 0.0538 |
| | 3.3984 | 0.0238 | 0.1106 | 2.2851 | 0.0520 |
| | 3.4000* | 0.0272 | 0.1104 | 2.2807 | 0.0546 |
| | | | $k = 5$ | | |
| 0.00 | 1.0000* | 1.0000 | 0.2252 | 4.9743 | 0.0460 |
| 0.09 | 2.0462 | 0.3866 | 0.2068 | 4.5168 | 0.0516 |
| | 2.0558 | 0.4449 | 0.2078 | 4.5414 | 0.0524 |
| | 2.0954 | 0.3589 | 0.1970 | 4.2774 | 0.0514 |
| | 2.1946 | 0.5221 | 0.2112 | 4.6252 | 0.0486 |
| | 2.2000* | 0.5282 | 0.2124 | 4.6549 | 0.0508 |
| 0.25 | 2.7652 | 0.0081 | 0.1648 | 3.5109 | 0.0486 |
| | 2.8100 | 0.0768 | 0.1800 | 3.8689 | 0.0556 |
| | 2.9094 | 0.0112 | 0.1728 | 3.6985 | 0.0600 |
| | 2.9914 | 0.1745 | 0.1758 | 3.7693 | 0.0516 |
| | 3.0000* | 0.1875 | 0.1818 | 3.9118 | 0.0628 |
| 0.64 | 4.1957 | 0.0039 | 0.1250 | 2.6033 | 0.0560 |
| | 4.2000* | 0.0067 | 0.1184 | 2.4568 | 0.0598 |
| 0.98 | 4.9600 | 0.0001 | 0.0610 | 1.2271 | 0.0328 |

*Note.* * indicates that all correlations in **P** are equal.

*Table 3.*
Expected values for the average $r^2$, maximum eigenvalue ($\lambda_{max}$) and determinant (|P|) with empirical Type I error rate, estimated $c$ from (14) and corrected Type I error rate (11) for $k$ = 7 and 10 predictors.

| $E(\bar{r}^2)$ | $E(\lambda_{max})$ | E( \|P\| ) | $k = 7$ $p$-value | $c$ (14) | Corrected $p$-value |
|---|---|---|---|---|---|
| 0.000 | 1.0000* | 1.0000000 | 0.3024 | 7.0206 | 0.0476 |
| 0.090 | 2.3742 | 0.1097200 | 0.2664 | 6.0396 | 0.0548 |
| | 2.5193 | 0.1764100 | 0.2708 | 6.1569 | 0.0570 |
| | 2.5539 | 0.2311600 | 0.2738 | 6.2373 | 0.0546 |
| | 2.7745 | 0.3024000 | 0.2720 | 6.1890 | 0.0596 |
| | 2.8000* | 0.3294200 | 0.2602 | 5.8755 | 0.0576 |
| 0.250 | 3.3159 | 0.0016268 | 0.2138 | 4.6896 | 0.0544 |
| | 3.5875 | 0.0004800 | 0.2040 | 4.4481 | 0.0486 |
| | 3.6922 | 0.0125600 | 0.2234 | 4.9291 | 0.0564 |
| | 3.7627 | 0.0001200 | 0.2072 | 4.5266 | 0.0540 |
| | 3.9779 | 0.0554100 | 0.2344 | 5.2072 | 0.0660 |
| | 4.0000* | 0.0625000 | 0.2232 | 4.9241 | 0.0558 |
| 0.640 | 5.7938 | 0.0002600 | 0.1492 | 3.1501 | 0.0586 |
| | 5.8000* | 0.0003700 | 0.1466 | 3.0906 | 0.0574 |
| | | | $k = 10$ | | |
| 0.000 | 1.0000* | 1.0000000 | 0.3920 | 9.7007 | 0.0456 |
| 0.153 | 2.3770* | 0.5333101 | 0.3960 | 9.8289 | 0.0512 |
| | 4.0039 | 0.0003575 | 0.2986 | 6.9147 | 0.0516 |
| 0.187 | 2.6830* | 0.4163279 | 0.3634 | 8.8045 | 0.0512 |
| | 4.4190 | 0.0000643 | 0.2794 | 6.3882 | 0.0514 |
| 0.205 | 2.8450* | O.3608996 | 0.3636 | 8.8107 | 0.0484 |
| | 4.7727 | 0.0000260 | 0.2900 | 6.6771 | 0.0532 |
| 0.327 | 3.9430* | 0.1116766 | 0.3474 | 8.3206 | 0.0624 |
| | 5.8601 | 0.0000005 | 0.2462 | 5.51004 | 0.0548 |
| 0.400 | 4.6000* | 0.0463574 | 0.3062 | 7.12708 | 0.0558 |
| | 6.5298 | $4.49^{e-8}$ | 0.2268 | 5.01464 | 0.0536 |
| 0.532 | 5.7880* | 0.0062336 | 0.2784 | 6.36115 | 0.0614 |
| | 7.4954 | $2.84^{e-9}$ | 0.1840 | 3.96428 | 0.0550 |

*Note.* * indicates that all correlations in P are equal.

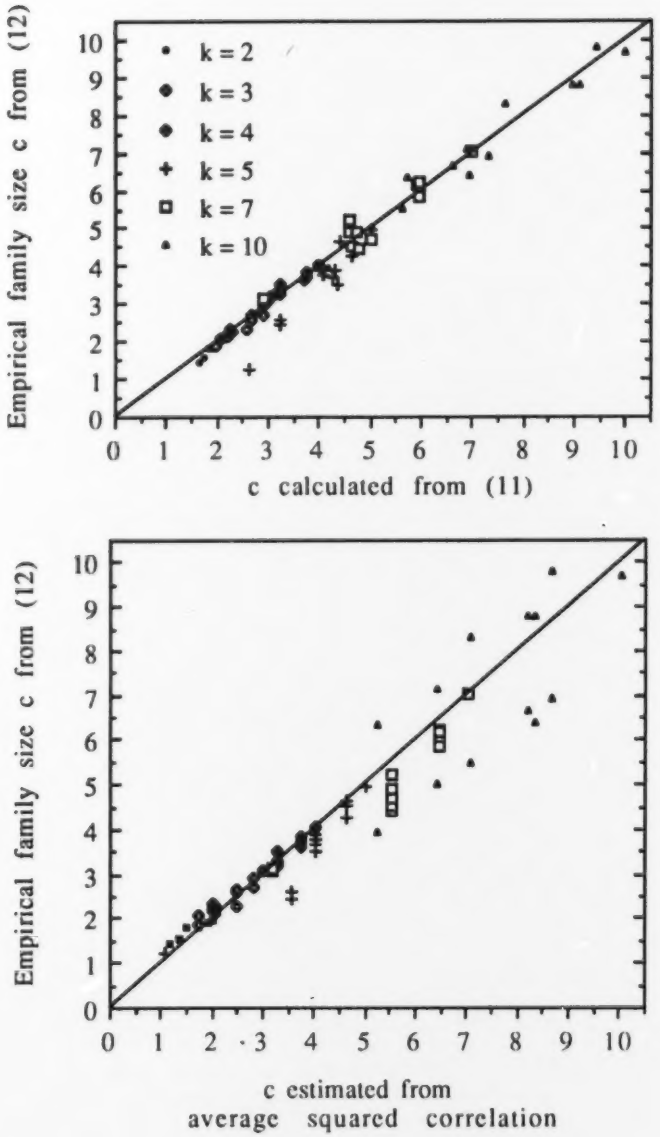


Figure 1. Empirical family size, $c$, derived from (12) as a function of the estimated family size from (11; upper panel) and from the average squared correlation (lower panel).